MACHINE LEARNING IN CYBERSECURITY

Network Intrusion Detection System

PROBLEM STATEMENT

Cybersecurity in 2020

- 64% of companies have experienced web-based attacks
- Small businesses target 43% of time
- Average data breach to cost $150 billion
- About six months to notice

The Dataset

- DARPA Intrusion Detection Evaluation Program (1998)
- MIT Lincoln Labs
- Simulated 494k connections (benign and variety of attacks) in military network environment

Attack Categories

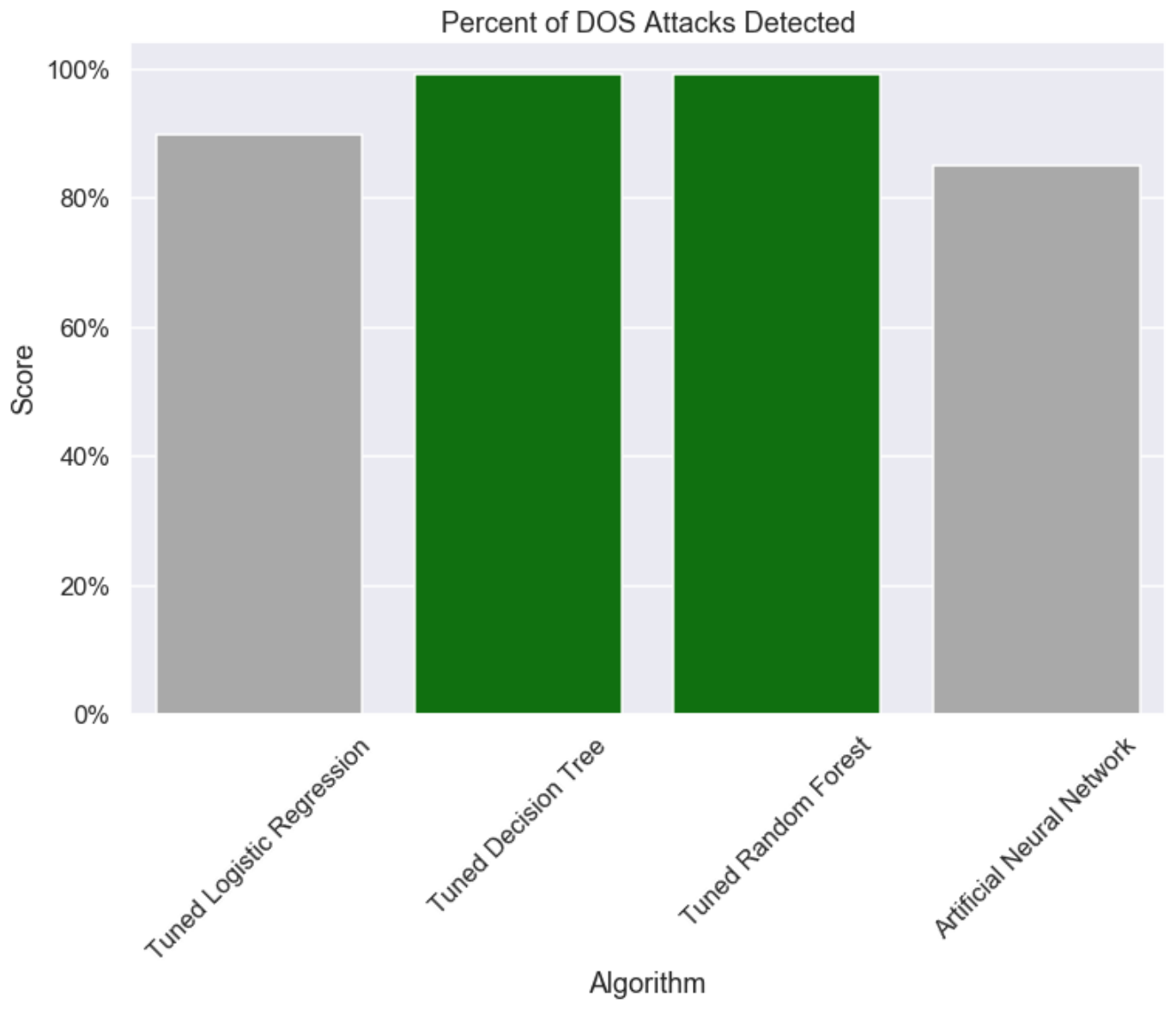
- Denial of Service (DOS)
- User-to-Root (U2R)
- Probe
- Remote-to-Local (R2L)

Intrusion Detection System

- Train various classifier on 494k connections containing 24 unique attack types
- Test classifiers on 311k connections containing 38 unique attack types

FINAL MODEL PERFORMANCE ON TEST SET

PERFORMANCE DETECTING DOS ATTACKS

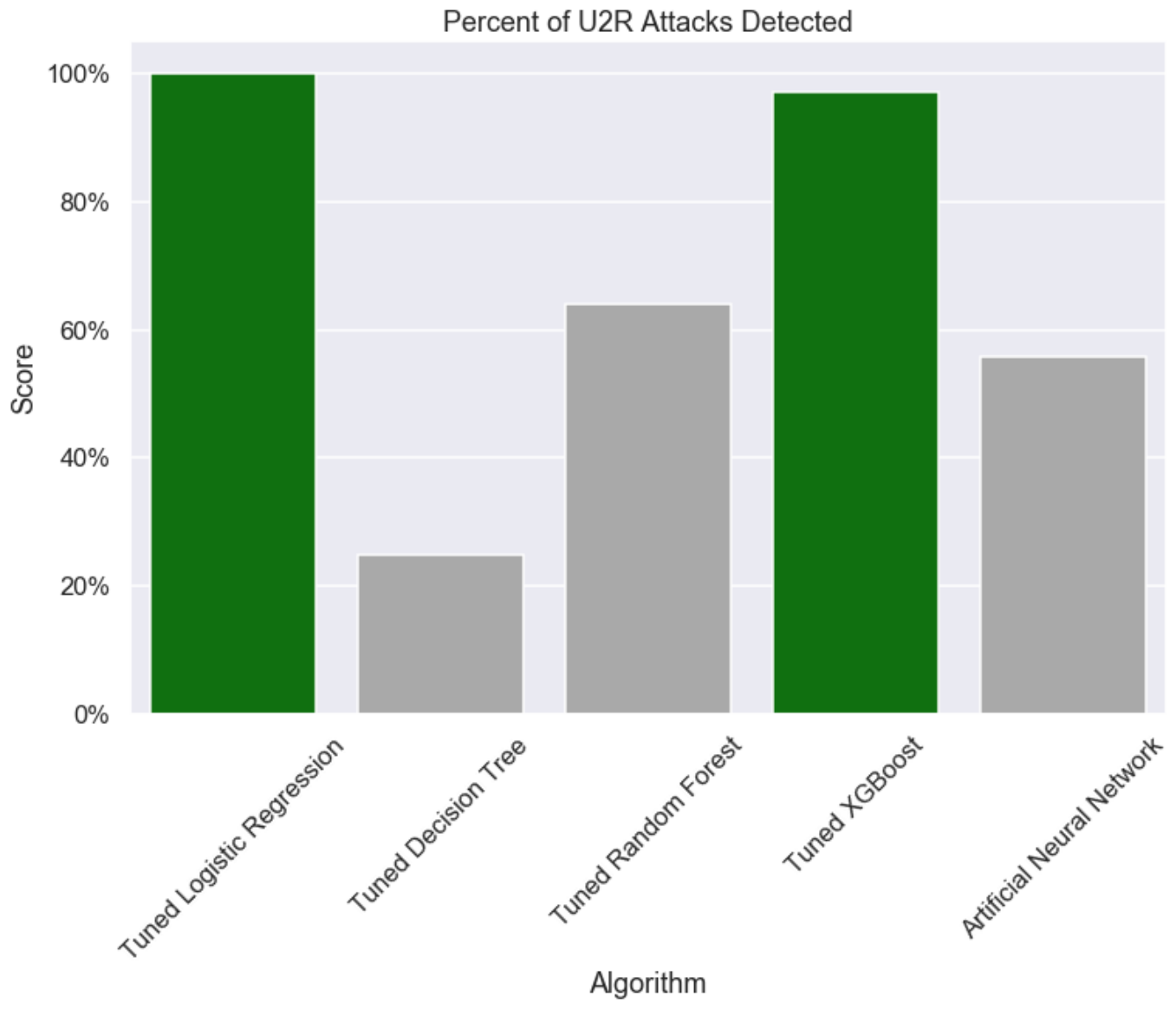


Attacks Detected

Decision Tree: 99.2%

Random Forest: 99.2%

PERFORMANCE DETECTING U2R ATTACKS

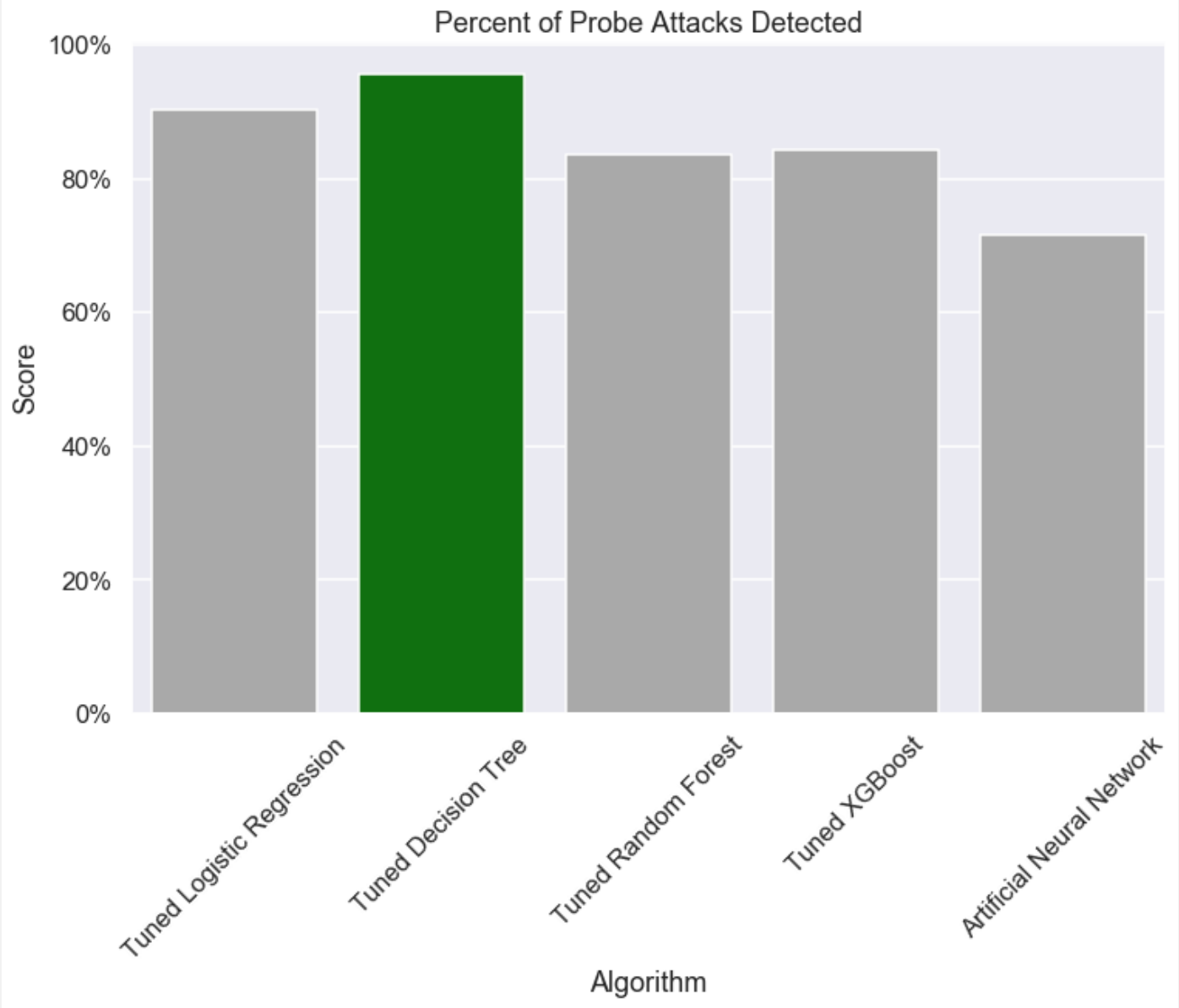


Attacks Detected

Logistic Regression: 100%

XGBoost: 97.1%

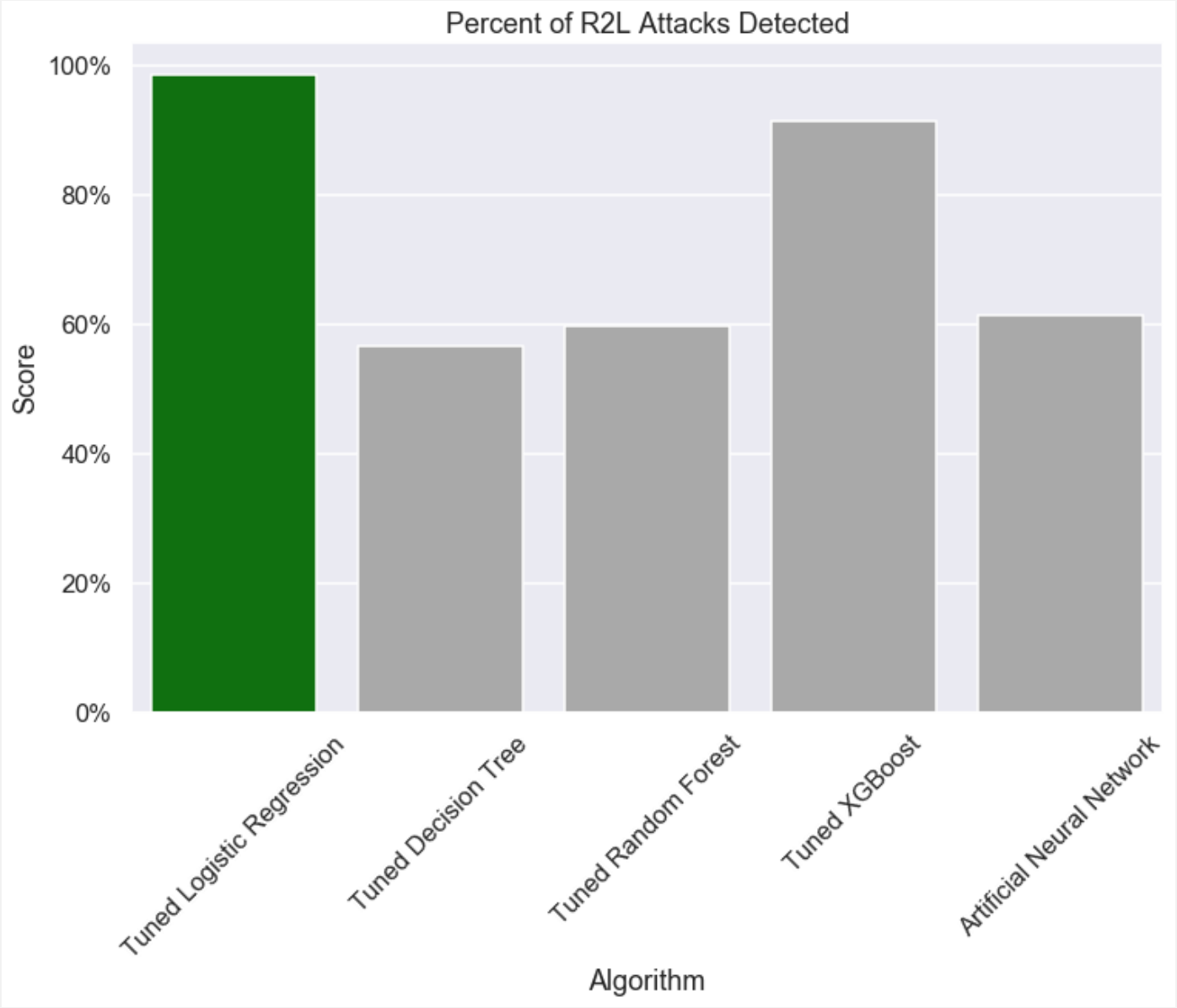
PERFORMANCE DETECTING PROBE ATTACKS



Attacks Detected

Decision Tree : 95.7%

PERFORMANCE DETECTING R2L ATTACKS



Attacks Detected

Logistic Regression: 99.2%

FUTURE WORK

Employ Rigorous Feature Selection

- Many unimportant features could be hurting model accuracy
- Employ advanced techniques to determine salient features

Address Attack Category Class Imbalance

- Most connections in training data were normal or DOS attacks (99%)
- Address attack category imbalance or get more data

Spend More Time on Model Tuning

- Specifically XGBoost and artificial neural network classifiers

Test Anomaly Detection Methods

- Train model exclusively on “normal” connections
- Test model to see if it can recognize abnormal connections (i.e. attacks)

THANK YOU

Braydon Charles Janecek

✉ *braydoncharlesjanecek@gmail.com*

🔗 *https://bcjanecek.github.io/*